Dealing with hierarchically clustered data

Missing value analyses and imputations

Working Paper

Author(s):
Kowald, M.; Arentze, T.

Publication date:
2010-11

Permanent link:
https://doi.org/10.3929/ethz-a-006253157

Originally published in:
Arbeitsberichte Verkehrs- und Raumplanung 656

Dealing with hierarchically clustered data: Missing value analyses and imputations

M Kowald (ETH Zurich)
T Arentze (Eindhoven University of Technology)

Arbeitsberichte Verkehrs- und Raumplanung 656 November 2010

Institut für Verkehrsplanung und Transportsysteme
Institute for Transport Planning and Systems

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

1. Introduction

In a joint survey project ETH Zurich and TU Berlin collect information on iteratively connected personal networks. Using an ascending sampling strategy called snowball sampling aims to ask respondents for names and addresses of their social contacts and continue the data collection with these contacts by again asking them to report their social contacts and so on. Snowball samples allow to focus on the global structure of connected rather than isolated personal networks and, with that, capture network properties behind isolated persons' horizons (for details on the survey methodology and instrument see Kowald et al., 2009).
Whilst having the advantage to suvey information behind isolated respondents' horizons, snowball sampling also has disadvantages. First of all there are several sources for bias (an overview and a discription of measures to avoid or reduce bias is given in Kowald and Axhausen, 2010). In addition, snowball sampling requires the necessity of asking respondents not only for personal characteristics, but also for names and addresses of their social contacts to allow the snowball chain to continue. In combination with further questions on e.g. sociodemographics of respondents and their social contacts, this results in a high amount of response burden.
The present chapter introduces the data in terms of item non-response behavior and missing values. After discussing where and why missing values ocurred in the present project, an imputation strategy is presented. This strategy considers the complex data structure resulting from snowball sampling. Where possible data are imputed by focussing exclusively on the relevant egocentric network. Only where this is not possible the overall values of the whole sample are used.

2. Missing values and item non-response

The present study collects data on personal networks by using a name generator. This part of the survey instruments implies question to focus respondents, called egos, on that part of their network being of importance for the project, the alters. In addition, the name generators' questions enforce an ego's entire memory process and help to collect the network as completely as possible (Marin and Hampton, 2007). In this study the name generator focusses on leisure and in addition on emotionally important contacts.
Usually the name generator approach is used to focus on randomly found respondents and their social contacts. Given a large sampling frame the chance of

personal network overlaps, in terms of two egos sharing one alter, is small. Observing such overlaps in significant numbers allows inferences on the network structure behind isolated egos' horizons and is the aim of the present study. The social coherence of connected personal networks is an important objective when e.g. aiming to implement social networks in agent based microsimilations. To collect information on iteratively connected social networks the study combines the name generator technique with the snowball sampling methodology (for more details on the implementation of the snowball methdology in this survey study see Kowald and Axhausen, 2010).

Surveying egocentric network structures with the snowball technique implies the necessity of not only asking egos to report names of social contacts but also to ask them for their alters' postal addresses. Detailed address information are also needed to geocode home locations and calculate spatial distances between social contacts.

Furthermore, the project's aim of allowing inferences on the population's coherence makes it necessary to not exclusively focus on ego – alter relations but, in addition, to collect information on alter – alter relations as well. This is done with a sociogram, asking egos to report groups of alters from the name generator that make plans to spend free time together (for details on the sociogram see Kowald et al., 2009).

In summary, the questionnaire is devided into four parts. First, it asks for egos' characteristics and mobility biographies. Second, the name generator asks for leisure and emotionnaly important contacts. It provides space for 40 names. Egos who want to mention more names are asked to write them on an extra cheat of paper. Next, egos are asked to specify characteristics of each alter mentioned in the name generator as well as modes and contact frequencies used to maintain the relationship. This third part of the survey instrument is called a name interpreter. The last part is the sociogram, collecting information on alter – alter relations.

A response rate of around 26% is achieved (calculated according the COOP4 cooperation rate due to the suggestions of the American Association for Public Oppinion Research (AAPOR, 2009)) and considered as highly satisfying given the response burden and the level of confidentiallity of the questions (for more details on the survey instrument and surveying methodology see Kowald et al., 2009; Kowald and Axhausen, 2010).

The shares of missing values per question cover a wide rage. They differ between questions and parts of the questionnaire. Whilst the survey instrument contains 43 questions in total, the following missing value analysis only employs 37 questions.

Some questions were combined to get useful answers, e.g. whether or not an alter is of emotional importance for an ego or when ego is asked to specify contact frequencies, for other questions it is not possible to define whether or not a missing value corresponds to item non-response behavior or is the answer of a person who does not match an attribute of question. Although literature on survey instruments recommends a design allowing this distinction (e.g. Dillman, 2000) we did not use such additional answer opportunities to keep response burden lower and the survey instrument shorter. Table 1 gives an overview on the questions being considered in the missing data analysis and their amount of response burden which is calculated with a rating system from commerical survey research (for details see Axhausen and Weis, 2010). The share of missing values for part 3, the name interpreter is given as the average share of each ego's missing value shares. The 'combination'-column indicates if a variable is a combination of two or more original questions. A variable marked as 'filter' depends on an earlier question. In cases where an ego did not match the filter-criterion, these questions were not asked and these cases are not considered for the analysis of missing values.

Clearly both, question position and amount of response burden, influence the share of missing values. Non-response behavior becomes more frequent in case of difficult questions like egos' mobility biographies (Question 20 and 21) and increases the later a question is positioned in the survey instrument (Questions 24 to 36). A surprise is the relative high share of missing values in the variable collecting data on the number of working household members (question 10). We did not find an explanation for this yet. High shares of missing values can also be observed in egos' level of education (question 12) and when asking for alters' level of education (question 28). This can on the one hand be a result of the predetermined answer categories for this questions in combination with the snowball methodology. Because the snowball chain was not limited to a certain geographical area it sometimes left Switzerland to be continued in another country. This was expected and the answer categories for e.g. education level were formulated in a general way. However, achievements in education differ between countries. So, there may be cases were the answer categories resulted in confusion and item-non-response behavior. On the other hand it may be hard for some long duration relationships to remember an alter's educational background. This could, at least for the name interpreter be a partial explanation for the missing values.

Using a Spearman's test to check response burden and question position for their correlation with the share of missing values showed the relation between missing values and response burden being weakly positive, 0.381, and significant (p =

0.020). The relation between missing values and question position is highly positive, 0.853, and significant (p = 0.000).

Table 1 Questions considered in the missing value analysis

Nr.	Aim of the question	Combination	Filter	Amount of response burden [abs]	Share of missing values [%]
Ego's characteristics					
1	Drivers license	-	-	1	3.54
2	Car availability	-	x	2	3.72
3	Year of birth	-	-	2	0.35
4	Gender	-	-	2	0.88
5	Civil status	-	-	2	0.35
6	Country of birth	-	-	2	0.18
7	1. Citizenship	-	-	2	0.88
8	2. Citizenship	-	-	2	0.0
9	Number of household-members	-	-	5	0.0
10	Number of workers in household	-	-	1	3.72
11	Education duration	-	-	2	1.42
12	Level of education	-	-	3	6.55
13	Work status	-	-	3.5	0.53
14	Self employed/employed	-	x	1	0.53
15	Weekly work time	-	x	1	1.59
16	At work place or home	-	x	2	1.06
17	Work Internet	-	x	1	1.59
18	Private mobile	-	-	1	1.95
19	Private Internet	-	-	3	3.89
20	Biography: Number of first residents	-	-	20	6.02
21	Biography: Number of education places	-	-	20	12.04
22	HH-income	-	-	3	3.71

Table 1 Continued

Nr.	Aim of the question	Combination	Filter	Amount of response burden [abs]	Share of missing values [%]
Name generator					
23	Name generator	x	-	13	6.02
Name interpreter					
24	Gender	-	-	2	5.05
25	Civil status	-	-	2	5.99
26	Year of birth	-	-	1	10.42
27	1. Citizenship	-	-	2	6.05
28	Level of education	-	-	3	11.65
29	Kind of relationship	-	-	3	5.89
30	Relationship duration	-	-	1	15.01
31	Context of 1st meeting	-	-	3	12.82
32	Annual contacts face-to-face	x	-	4	12.76
33	Annual contacts phone	x	-	4	14.21
34	Annual contacts E-mail	x	-	4	16.40
35	Annual contacts online messaging	x	-	4	19.36
36	Emotionally important contact	x	-	2	13.53
Sociogram					
37	Sociogram	-	-	11	20.71

Combination = The variable is a combination of originally two or more questions;
Filter = Weather the question is relevant depends on a former question

In part 3, the name interpreter, the egos are asked all questions for each alter mentioned in the name generator. Therefore the amount of response burden has to be multiplied with the number of alters in a personal network. To check whether there are differences in item non-response behaviour due to the number of alters mentioned in the name generator, figure 1 shows the shares of missing values per questions for egos with 1-10 alters (top graph), with 11-20 alters (2nd graph), 21-30 alters (3rd graph) and egos mentioning 31 to 40 alters (bottom graph). It also shows

an moving average calculated over five consecutive variables. The vertical dotted lines represent the bondaries between different parts of the questionnaire.

The graph clearly shows that the instrument contains questions which result in very high amounts of missing values, independent from the number of alters in an egocentric network. Egos' and alters level of education (question 12 and 28) have been mentioned before. Also problematically is the variable for ego's and alter's relation duration (question 30). In terms of response burden it is considered as a simple numerical answer but it may be hard to remember the year of a first meeting for some relationships, especially older ones. In addition, the comparison between egos with different network sizes shows question 30 as being a benchmark characterized by an rapidly increasing share of missing values. Whether or not the share of missing values at later questions would be lower if this question was removed from the survey instrument or asked last cannot be tested. It would be worth trying in future studies as the share of missing values remains high after this point.

Increasing missing value shares can also be observed for the communication frequencies by different modes (questions 32 to 35). Each of these variables is calculated as the annual communication frequencies per mode by using two questions from the original survey instruments. Respondents were asked to mention the frequency and specify the unit that best fits from a predetermined list: per day, per week, per month, or per year. This results in an increased amount of response burden. In addition, some participants may have lost interest when answering these questions on communication behaviour consecutively. Increasing non-response is also present in the first part of the instrument when two consecutive questions are similar to each other like question 20, asking for the addresses of an ego's first residents in the live course and question 21, asking for education places over the live course.

Comparing the shares of missing values for different network sizes shows a surprising result. Intiutively it could be assumed that non-response behavior increases due to fatigue effects with an increasing network size as each question of the name interpreter is asked for each alter mentioned in the name generator. This tendency can clearly be found in the data as the moving average increases with each additional question. But there is also another tendency, which seems to be related to respondents motivation and patience.

Those mentioning small networks with between 0 and 10 alters (top graph) are weakly motivated. These 148 egos start with a high average share of missing values (4%) and this value increases with every further question. Especially in cases with

high amounts of response burden these respondents reacted with item-non-response. They finish the questionnaire with around 19% missing values on average.

Less variance can be observed for the 173 respondents with network sizes between 11 and 20 alter (second graph). They start with around 2% missing values and end with around 14% on average. It is worth noticing that this category is the densest in terms of respondents and contains the average network size not only of the present project but also of survey studies asking for similar kinds of information (see e.g. Carrasco, 2006; Frei and Axhausen, 2007). Overall egos within this categoy seem to have the lowest shares of missing values in the name interpreter and can therefore be labeled as highly motivated to answer the questions completely.

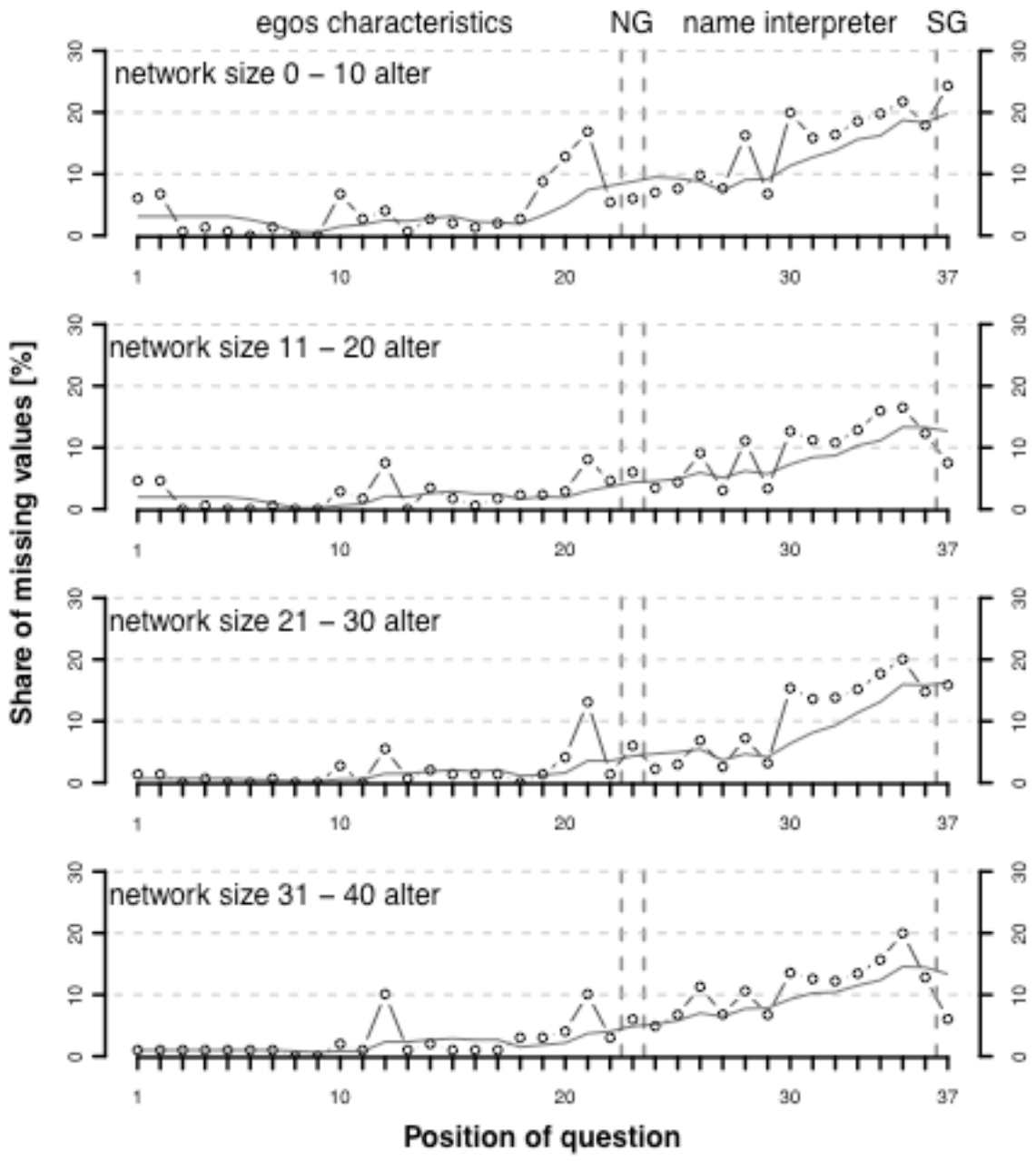
The third category of network sizes contains 145 egos (third graph) which are also motivated. Starting with a low share of missing values (around 1%) the non-response behavior of this group stays low egos' own characteristics. It increases rapidly at question 31 and increases even further in case of the questions on communication mades and frequencies. It seems that this group looses interest in filling out the questionnaire at some point. Alters and questions lying behind this point have a high probability of remaining unanswered. Because of their low shares of missing values in the beginning of the name interpreter, the behavior of this group can be described as motivated but fatiguing.

Contarily, the fourth group of only 99 egos with large network sizes (graph four) can be described as idealistic. They start with low missing value shares, on average 1%, and stop with relatively low ones, around 7%. Although the correlation between a question's position in the survey instrument and the response behavior is also present for this group they show a relative stable preference to answer the questions. Considering the high amount of responde burden resulting from the high number of reported social contacts this demonstrated a strong will to answer all questions.

In summary, a question's position and its amount of resonse burden are of significant influence on partipants' response behavior. Questions which are difficult to answer, whether because they are complex or demand difficult recall tasks because they deal with historical events should be avoided. In addition questions that are very similar to each other should be avoided. Furthermore, the present study implies effects which result from the answer categories. Of high importance is the motivation of the respondent. If it is high this seems to be the best garanty for a low item-non-response behavior.These findings are consistent

with literature on survey research (e.g. see Dillman, 2000) and behavioral theories (see Kowald, 2010).

Figure 1 Shares of missing values for egos with different personal network sizes (from top to bottom: 1) 1-10 alter n = 148; 2) 11-20 alter, n = 173; 3) 21-30 alter, n = 145; 4) 31-40 alter, n = 99)



3. A strategy for imputing missing values

Imputations aim to reduce the influence of missing values. This can be necessary when aiming to use statistical methods which require complete-case analysis like e.g. most types of regression analysis do. These methods methods exclude alll observations with missing values in either dependent or independent variables from the analyses. If an ordinary linear regression model would include many variables with missing values, it may occur that in the end only few complete observations are useable for estimation.

There are several methods for missing value imputations with a range from easy ones, like a mean imputation, to complex ones using modelling techniques to consider the relations between different variables in the imputation (for an overview on imputations and dealing with missing values see: Little and Rubin, 1987; Rubin, 1987). Which approach is choosen depends on the data that should be imputed and the statistical methods the data will be used for.

The present imputation approach aims to replace missing values in egos and alters sociodemographics to use the data in regression analyses. The method follows a 'best guess'-approach using monte carlo simulations. In this way the imputed variables' variances should be maintained. As the share of missing values is small for most variables, mostly under 10%, we think this approach is appropriate.

Data on personal networks are structured hierarchically in terms of egos reporting nests of alters. This structure can be used for imputations in characteristics which were surveyed for both egos and alters. Relevant variables are: gender, civil status, level of education and age.

The degree of similarity between an ego and her alters are used as the basis for imputations. In terms of network analysis this similarity is called status homophily. If e.g. an ego was a female and reported 12 alters, of which 7 were females while 3 were males and the gender attribute of two alters was missing, the degree of gender homophily for this egocentric network would be 70%. This threshold is used for the gender imputation. To do so, a random number between 1 and 100 must be drawn. If this random number was smaller 70 the imputed value would be 'female' and otherwise 'male'. The process is repeated for each missing gender attribute. For variables with more than two attributes the method uses the summed up shares of each attribute within a variable.

The method becomes more complex in terms of continuous variables. To impute missing values in age five age classes were defined for both egos and alters which use the quantiles for alters age and an additional category for persons over 60

years. This aims to have an extra class for retirees. The quantiles use alters' age because here more observations are avialable than for egos' age. However, the interquantile categories fit well for both age distributions. The first class is defined as persons junger than 37 years. The second class contains egos and alters between 38 and 45 years. Next is a class between 46 and 52 years, followed by a category for persons between 53 and 60 years. A missing value is imputed by adding ego's age with the average age difference between ego's and alters' within the ego's age category and a multiplication term of the standard deviation of egos' and alters' age difference within the age category and a randomly number drawn from the standard normal distribution. The multiplication term aims to add a positive or negative sign to the standard deviation and with that keep a certain amount of variance within the age categories.

In summary, this imputation method uses the degree of status homophily for each single egocentric networks to replace missing values. It follows a Monte Carlo simulation approach. Initially the method is used to impute missing ego values with alters characteristics. In a second step the alter values are imputed by using egos' characteristics. Table 2 provides an overview on the imputation results. Large numbers of missing valued were imputed. In case of some variables all missings were replaced. Missings remained in cases where no information on both egos' and the alters' characteristics were available. Because egos mention different numbers of alters imputations within each nest can have different weights for the overall variable distribution. However, all variables were tested whether their distribution changed significantly due to the imputiation. None of these tests showed a significant result.

Table 2 Comparing the distributions of variables before and after imputations

Variable	Missings before imputation [abs]	Share of replaced missing values [%]	Difference in distribution (p-values)
Egos' characteristics			
Egos' gender	83	100.0	0.612
Egos' civil status	36	100.0	1.000
Egos' education	800	99.9	0.417
Egos' age	36	100.0	0.956*
Alters' characteristics			
Alters' gender	408	69.4	0.810
Alters' civil status	534	65.4	0.981
Alters' education	1077	78.8	0.906
Alters' age	991	100.0	0.303*

N egos: 565; N alter: 10'800

(*) = Tested with a t-test; all other characteristics tested with a χ^2-independence test.

Information on both egs and alters and with that on status homophily is only available for the minority of variables. In addition, there are still missing values in some of these characteristics. In a second step, missing values of variables which are only available for either egos or alters are imputed. Also missings remaining in gender, civil status, education level or age are imputed. The method again follows a 'best guess' approach and uses Monte Carlo simulations. But instead of using the degree of status homophily the imputation depends on the attribute distribution within a variable. Relevant variables for egos are: level of education, duration of education, number of first residents in course of live, number of education places in course of live, driver's license, car availability, work status, terms of employment, work place, amount of work, Internet availability at home and at work, mobile telefone, number of persons in the household, and household income. Work related variables were only imputed for egos that are working.

Relevant variables of alter characteristics are: gender, civil status, education, kind of relation with ego, duration of their relation, context of first meeting between alter and ego and whether alter is an emotionally important contact for ego.

In case of a missing gender attribute of an alter the relative gender shares over all alters are used as the basis for the Monte Carlo simulation. In this way the variance within a variable is used as the basis to replace missing values. In case of continuous variables quantiles were calculated and used to achieve variance in the imputations. In case of age the method again uses the five age classes defined above. A first Monte Carlo simulation uses the shares of alters (or egos) within the categories to choose a class. The mean age within this class is added to average difference between the mean and alters'-ages in the class and an multiplication term of the intraclass standard eviation and a randomly drawn number from the standard normal distribution. In some cases these imputations for continuous variables used logical boundaries. E.g. the length of a duration between an ego's and alter's was limited to the age of the younger persons, either ego or alter. Egos' age values had to be over 18 years as younger persons were not recruited. Egos' and alters' age were limited to 90 years, the highest age reported in the survey.

In summary, all missing values in the relevant variables were replaced. The variables' distributions did not change significantly. In cases with high numbers of missing values also multivariate imputations were tested. In e.g. the number of education places in course of an ego's live we tried to use the length of education in years as a proxy for the number of education places. Because the distributions within different lengths-classes were very similar we decided to keep the imputation approach simple.

The variables dealing with egos' and alters' contact modes and frequencies stay unimputed. Although they have high shares of missing values it can be assumed that these missings are due to contact modes which ego and alter do not use to stay in contact. In cases where contact frequencies for all modes are missing the values are not replaced. In cases where frequencies for at least one contact mode are reported, missing values for the other modes are replaced by logical zeros. In other words, it is assumed that there are zero annual contact between an ego and an alter by this mode.

All imputed variables are added to the data. In case of analyses the unimputed as well as the imputed variables can be used. This way of data storage allows for flexibility in deciding whether imputed or unimputed values should be used in a specific kind of analysis.

4. Conclusion

This working paper introduced data from a snowball survey on personal leisure networks in terms of item non-response-behaviour and missing values. Although the survey involves a high amount of response burden the share of missing values is low for most variables. Clearly item-non-response behaviour is positively correlated to the position in which a question is asked in the survey instrument and the burden a question implies. Questions, which are similar to each other result in an increasing share of missing values when they are asked consecutively. In addition, a high share of missing values was found for questions, which need high recall efforts to be answered like historical information. Because in the present study the amount of response burden is related with the number of reported social contacts a detailed missing value analysis was done by differentiating between four classes of network size. This showed that respondents with small networks (<= 10 social contacts) have a stronger tendency to item-non-response behaviour than those with larger networks. In other words, respondents who report many social contacts are motivated to participate. They try to answer the questions although the increasing response burden leads to fatigue effects at least for some respondents. Finally, respondents with large networks (>30 social contacts) can be labelled as idealistic as they show only little item-non-response-behaviour although response burden is very high for them.

The second topic of the paper was to introduce a strategy to impute missing values by considering the sampling methodology. Snowball samples result in hierarchical data structures as each respondent reports several social contacts. The present imputation strategy aims to use this structure. Following a 'best guess'-approach it uses the degree of homophily, the similarity between egos and alters in certain characteristics, as a threshold for Monte Carlo simulations. Accordingly to the thresholds missing values are imputed for egos' and alters' characteristics. It was shown that this strategy is effective and avoids changes within variables' distributions. For characteristics where no similarity threshold is available the variable's distribution over all cases is used as the basis for Monte Carlo simulations. This imputation approach is rather simple than complicated and is mostly used for egos' and alters' socio-demographics. It aims to avoid loosing whole observations in case of statistical complete-case analysis like linear regression models.

5. Acknowledgements

The author acknoledges the help of Prof. Theo Arentze from Eindhoven University od Technology who helped a lot in the development of the imputation strategy. He was always willed to share his thoughts and give a helping hand when questions occured.

6. Literature

The American Association for Public Oppinion Research (2009).

Axhausen, K.W. und C. Weis (2010) Predicting response rate: A natural experiment, *Survey Practice*, **3** (2), http://surveypractice.org/2010/04.

Carrasco, J.A. (2006) Social activity-travel behaviour: A personal network approach, *Dissertation*, University of Toronto, Toronto.

Dillman, D.A. (2000) *Mail and Internet Surveys. The Tailored Design Method*, Wiley & Sons, New York.

Frei, A. and K.W. Axhausen (2007) Size and structures of social network geographies, *Arbeitsberichte Verkehrs- und Raumplanung*, **444**, IVT, ETH Zurich, Zurich.

Kowald, M., A. Frei and J.K. Hackney, J. Illenberger and K.W. Axhausen (2009) Using an ascending sampling strategy to survey connected egocentric networks: A field work report on phase one of the survey, Arbeitsberichte Verkehrs- und Raumplanung, 582, IVT, ETH Zurich, Zurich.

Kowald, M. and K.W. Axhausen (2010) Spatial distribution of connected leisure networks: Selected results from a snowball sample, *Arbeitsberichte Verkehrs- und Raumplanung*, **614**, IVT, ETH Zürich, Zürich.

Little, R.J.A. and D.B. Rubin (1987) *Statistical analysis with missing data*, Wiley & Sons, New York.

Marin, A. and K.N. Hampton (2007) Simplifying the personal network name generator: Alternatives to traditional multiple and single name generators, *Field Methods*, **19** (2) 163 – 193.

Rubin, D.B. (1987) *Multiple imputation for nonresponse in surveys*, Wiley & Sons, New York.